Giving a Little Help to Girls? Evidence on Grade Discrimination and its Effect on Students Achievement

Camille Terrier

PARIS SCHOOL OF ECONOMICS
ECOLE D'ECONOMIE DE PARIS

WORKING PAPER N° 2014 – 36

Giving a Little Help to Girls? Evidence on Grade Discrimination and its Effect on Students Achievement

Camille Terrier

JEL Codes: I21, I24, J16

Keywords: Gender, grading, discrimination, progress

PARIS-JOURDAN SCIENCES ECONOMIQUES
48, BD JOURDAN – E.N.S. – 75014 PARIS
TÉL. : 33(0) 1 43 13 63 00 – FAX : 33 (0) 1 43 13 63 10
www.pse.ens.fr

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE – ECOLE DES HAUTES ETUDES EN SCIENCES SOCIALES
ÉCOLE DES PONTS PARISTECH – ECOLE NORMALE SUPÉRIEURE – INSTITUT NATIONAL DE LA RECHERCHE AGRONOMIQU

Giving a Little Help to Girls? Evidence on Grade Discrimination and its Effect on Students Achievement *

Camille Terrier†

Paris School of Economics

This version: October 29, 2014

Abstract

This paper tests whether we observe sex-discrimination in teachers' grades, and whether such biases affect pupils' achievement during the school year. I use a unique dataset containing standardized tests, teachers' attributed grades, and pupil's behavior, all three at different periods in time. Based on double-differences, the identification of the gender bias in teachers' grades suggests that (i) girls benefit from a substantive positive discrimination in math but not in French, (ii) girls' better behavior than boys, and their initial lower achievement in math do not explain much of this gender bias. Then, I use the heterogeneity in teachers' discriminatory behavior to show that classes in which teachers present a high degree of discrimination in favor of girls at the beginning of the year are also classes in which girls tend to progress more over the school year compared to boys.

*I would like to thank Francesco Avvisati, Marc Gurgand, Nina Guyon and Eric Maurin for sharing their dataset. This paper benefited from discussions and helpfull comments from Marc Gurgand, Eric Maurin, Corinne Prost, Thomas Breda, Thomas Piketty and Julien Grenet, as well as seminar participants from the Paris School of Economics, European Doctoral Program in Quantitative Economics jamboree, University Paris 1 - Pantheon Sorbonne and French Ministry of Education Workshop.

†Electronic address: camille@cterrier.com

1 Introduction

This paper refers to two puzzles regarding pupil's success. The first one is the persistent achievement gap which exists between boys and girls at the earliest stage of schooling. In most OECD countries, boys outperform girls in mathematics, whilst the opposite is observed in humanities (OECD, 2009). International comparative studies of educational achievement provide evidence of this gender gap. In the 2011 TIMSS assessment of mathematical knowledge of 4th grade pupils, of the 24 countries with a statistically significant gender difference, 20 had differences favoring boys – among which the United States, Finland, Norway, Austria, Korea, Germany and Italy. Regarding humanities, in nearly all of the 45 countries participating to the PIRLS assessment, 4th grade girls outperformed boys in the reading achievement in 2011.

The second puzzle relates to the evolution of this achievement gap over grades. In mathematics, achievement inequalities vanish when pupils grow up, whilst they persist in humanities. In the former subject, TIMSS assessments have shown gender differences in achievement to favor boys on average at the fourth grade, but to disappear or favor girls at the eighth grade, although the situation varies considerably from country to country. On the contrary, recent research in the United States finds that girls have an advantage in reading at all grades from kindergarten through the eighth grade (Robinson and Lubienski, 2011), and PISA 2009 reports that 15-year-old girls perform consistently better in reading than boys (OECD, 2009).

These puzzles raise two related questions: what explains the gender inequalities in achievement? Why does this gap vanish in math but persist in humanities? This paper sheds new light on the issue of grade discrimination - a potential explanation for gender inequalities - and provides evidence on the new question of the impact of gender biases on girls' progress relative to boys. Gender gaps in achievement and their persistence across grades are of particular concern since a small achievement gap might cause loss of self-confidence from low achievers, but also greater inequalities in tracks chosen, subjects of study at university, and finally in wages. In an effort to understand the origins of these gender inequalities, research has proven that teachers' stereotypes affect their pupils' success, notably because stereotypes can bias teachers' assessment and grades (Bar and Zussman 2012, Burgess and Greaves 2009). It has often been believed that teachers have negative stereotypes towards girls. Common stereotypes mentioned in prior literature consist in believing that girls are less competitive than boys, less logical, less adventurous and that they rely more on effort than on ability to succeed (Tiedemann 2000, Fennema and Peterson 1985, Fennema et al. 1990).

In the literature, the question of whether teachers' stereotypes towards girls affect their assessment has received quite a lot of attention. Yet empirical results differ on the direction of this potential discrimination. Some research concludes that girls benefit from discrimination in all subjects (Lindhal 2007, Lavy 2008, Robinson and Lubienski 2011, Falch and Naper 2013), while others find no gender discrimination (Hinnerich et al. 2011). Ouazad and Page (2012) and Dee (2007) observe that gender discrimination is not generalized but rather depends on teachers' gender, while Breda and Ly (2012) find that discrimination depends on the degree to

which the subject is "male-connoted". Besides the inconclusive nature of this existing literature, two concerns are frequent in prior studies. Most of them are not able to disentangle a pure gender bias from a discrimination related to pupils' behavior. Hence the risk of biased estimates due to omitted variables. Some studies also measure discrimination by comparing blind and non-blind scores, but this is not sufficient if both tests are not similar, and thus do not measure the same abilities. One contribution of this paper is to address both concerns.

Another key question is whether grade discrimination affects pupils' progress. Prior research has studied how teachers' expectations influence student achievement through self-fulfilling prophecies (Jussim and Eccles, 1992). As far as grades are a signal of teachers' expectations, they might predict pupils' achievement. Yet, this literature does not provide insights on the mechanisms through which grades could affect pupils' progress. Firstly, a positive bias might reduce the 'stereotype threat' effect. The latter arises when girls or minority groups perform poorly for the sole reason that they fear confirming the stereotype that their group performs poorly (Steele and Aronson 1995, Hoff and Pandey 2006). The apprehension it causes might disrupt women's math performance (Spencer et al. 1999). Therefore, over-grading girls can reduce their anxiety to be judged as poor performers when they undergo a math exam. Secondly, teacher-assigned grades have been proven to affect students' math self-concept and interest (Trautwein et al. 2006, Marsh and Craven, 1997), and therefore to potentially affect their achievement (Bonesronning, 2008). Finally, Mechtenberg (2009) provides a theoretical model of how biased grading at school can explain gender differences in achievements. School results are defined as a combination of talent and effort, the latter being the channel through which grade discrimination could affect future cognitive achievement. Although the link between biased grading and pupils' achievement has long been an important research question in education sciences, to my knowledge this is the first economic study to provide empirical evidence on how grade discrimination affect pupils' progress.

I use a rich student-level dataset produced by Avvisati et al. (2014). Three features make this dataset unique. Firstly it includes two different measures of a pupil's ability: a 'blind' score and a 'non-blind' score, which will enable me to identify the gender bias. 4519 pupils in 6th grade were required to take a standardized test at the beginning and at the end of the year. These tests are graded anonymously by an external corrector. They can be considered as blind scores free of any teachers' stereotypes. In addition to these blind scores, grades attributed by teachers were collected during the school year – hence non-blind and potentially affected by teachers' stereotypes. As long as both blind and non-blind scores measure the same skills, the blind score can be considered as the counterfactual measure to the non-blind score. A second advantage of this dataset is that it contains extensive information on pupils' behavior in the classroom. This allows me to disentangle grade favoritism related to gender from favoritism related to pupils' behavior. Thirdly, the blind score is available at the beginning and at the end of the school year. This gives me the unique opportunity to look at the impact of discrimination on pupils' progress.

I use a double-differences (DiD) strategy to identify the existence of gender biases in grades. Discrimination is defined as the average gap between non-blind and blind scores for girls, minus this same gap for boys. Prior research has used this method to estimate gender discrimination (Falch and Naper 2013, Breda and Ly 2012, Lavy 2008, Goldin and Rouse 2000, Blank 1991). Overall I find strong evidence for a substantial bias in favor of girls in math, representing 0.31 points of the s.d. No discrimination is observed in French. Controlling for pupils' punishment slightly decreases girls' positive bias in math so that a share of what I interpret as gender discrimination is actually a "good behavior bias". Similarly, controlling for pupils' achievement at the beginning of the year slightly decreases the gender bias in math, due to the fact that girls perform lower than boys in this subject, and that low performers tend to be favored by teachers. Finally, independently from any gender bias and academic achievement, I find suggestive evidence that being turbulent has a substantial negative effect on teachers' grades, and that being a low-achiever has a sizable positive effect on teachers' grades, consistent with an encouragement effect. These results are robust to a variety of alternative specifications that account for the fact that the blind and the non-blind scores might not measure the same abilities, that they are not filled in at the same date, and finally that girls might be more stressed than boys. My findings shed new light on gender bias in teachers' grades, in particular on the role of girls' behavior in teachers' gender bias. They tend to confirm existing studies which find that girls are favored by teachers in math (Falch and Naper 2013, Breda and Ly, 2012 ...).

In this subject, along with the bias observed in favor of girls, different patterns of progress between boys and girls are observed, raising the new and key question of the link between grade discrimination and pupils' subsequent progress. Therefore, I then focus my analysis on whether teachers' biases affect girls' progress over the school year. The identification strategy, based on class level data, exploits the high variation in teachers' discriminatory behavior: not all teachers favor girls, and among those who have a biased assessment of girls relative to boys, some are more biased than others. Taking advantage of both this heterogeneity and the quasi-random assignment of pupils to teachers who discriminate, the identification stems from a comparison of the relative progress of girls (as compared to boys) in classes where the teacher displays a high degree of discrimination, to the progress of girls in classes where the teacher does not discriminate much.

The key finding is that classes in which girls benefit from a high degree of positive discrimination (relative to boys) are also classes in which girls progress more (relative to boys). To my knowledge, this is the first study to examine empirically the impact of discrimination on pupils' progress, and to show that positively rewarding pupils has the potential to make them progress more. This result is consistent with two mechanisms mentioned in prior literature. Favoring girls can reduce the stereotype threat they suffer from, and hence reduce their apprehension when filling in a math exam. This could explain why biases affect girls' relative progress in math but not in French, a subject in which girls might suffer less from stereotypes threats. Positive biases can also affect girls' interest and self-confidence for math. However, my results tend to

challenge Mechtenberg's (2009) theoretical predictions according to which due to their awareness of receiving biased grading, girls would be reluctant to internalize good grades in math.

Taken together, these results build upon an important literature suggesting that teachers' grades are biased. My findings confirm the existence of such biases, but more importantly they highlight that gender discrimination can have long-lasting effects on girls' achievement relative to boys. I provide a new explanation for the fact that the achievement gap vanishes in math but persists in French. This is particularly relevant for the ongoing debate about policies aimed at promoting gender equality at school. Advocates of such policies usually focus their argumentation on the fact that teachers' grades can be a source of inequalities at school. My findings bring this argument one step further by highlighting that, over the long term, teachers' biases can also play a large and lasting role on the evolution of the gender inequalities at school.

The article proceeds as follows. Section 2 presents the dataset and gives some descriptive statistics. Section 3 defines a simple model of grade attribution, discusses the identification of gender discrimination in grades, and presents the results. Section 4 presents a model of pupils' progress, discusses the identification of the causal effect, and presents the results. Section 5 concludes.

2 Data

2.1 The dataset

I address the question of teachers' assessment bias by using a French dataset which contains 35 secondary schools, 191 classes, and 4519 pupils in 6th grade, hence 11 years old (Avvisati et al. 2014). The dataset provides two sources of information on pupils' achievements. Firstly, students have completed two standardized tests at the beginning and at the end of school year. These tests have been created by the French Education Ministry. They are identical across schools and classes and are externally graded. Every year, all French pupils used to take this national test at the beginning of 6th grade in order to assess their cognitive skills. Students also completed a similar test at the end of the school year, so that I have measures of pupils' abilities at two different periods. Knowledge on French and mathematics has been tested. The external correction of tests implies that the grader has no information on the gender, social background or school attended by pupils. Hence, these scores may safely be assumed to be free of any bias caused by stereotypes from an external examiner. The second source of information on children's achievements is provided by teachers' assessment of their own pupils. A pupil has a different teacher in each subject and all teachers report pupils' average grade on end-of-term report cards. In this study, I focus on mathematics and French grades given during the first and last term of the school year. In so far as teachers have permanent contacts with the pupils they teach, these average grades may reflect biases occasioned by teachers' gender stereotypes. Thus, I have two different scores meant at measuring students' knowledge. I use the term "blind scores" to describe test scores that have been anonymously graded. When grades have been

given by teachers who know pupils' gender and identity, I describe them as "non-blind scores".

It is worth mentioning that the standardized tests are high-stakes for neither the students nor the teachers. For students, they are a pure administrative evaluation aimed at reporting pupils' average achievement by schools to the Ministry. For teachers, their evaluations or salaries do not depend on their pupils' results to these tests so that they have no incentive to 'teach to the test'. The standardized tests are also taken in the same conditions as ordinary class exams: pupils fill in the test in their usual classroom and their teacher gives the instructions. Only the content of the tests differ, an issue that I will discuss further in the remaining of the paper.

A unique feature of this dataset is that it contains a rich set of measures of pupils' behavior for each of the three school terms. I have information on whether pupils were given an official "disciplinary warning", whether they were definitively excluded from the school, temporarily excluded from the school or from the class, whether they were put in detention or received blâmes[1]. Temporary exclusions signal violent behaviors or repeated transgression of the rules. They are sentenced by the school head. All these sanctions can be cumulated by pupils. Finally, the dataset contains administrative information on children: gender, parent's profession, grade retention, birth date... The schools included in this dataset are mostly located in deprived areas. Therefore they are not representative of all French pupils, an issue that I will discuss in a further section.

2.2 Sample Size and Balance check of attrition

The dataset contains 4519 pupils. For 555 of them (12.3%), one or more test score is missing during first term so that the sample is unbalanced. Missing scores might be blind or non-blind scores, in Math or French. The sample of pupils with no missing grades in Math and French contains 3964 observations – 4068 in Math only and 4058 in French. In order to test if pupils with one or more missing variable are different from those with no missing variables, I implement a balance check and compare several characteristics across both groups of pupils. I call the sample of 3964 pupils for which all test scores are available the "reduced sample" (as opposed to the full sample which includes the 4519 observations). The sample of 555 pupils for which at least one test score is missing is called the "missing sample". Results are presented in table 1.

[TABLE 1 HERE]

Pupils for which one or more test score is missing have different characteristics from pupils with no variable missing. They have systematically lower test scores in both blind and non-blind scores. For instance, in French during first term, their blind score is on average 0.453 points lower. There are also 9.7 percentage points fewer girls in the sample with missing variables. Parents belong less to the Upper or Lower class. Hence, we can expect parents being more middle class. Finally, regarding punishments, we do not observe important differences.

[1] Blâmes are official warning given by the school's administration when a pupil behaves badly in a repeated way.

Considering these differences, analyzing discrimination with the sole balanced sample is not satisfactory. Although this sample allows comparing results obtained with the same subset of pupils, it might yield results that suffer from a selection bias, hence being non-representative of the whole sample. In the remaining of the paper, I systematically run regressions on both samples: the sample of 3964 observations with no missing variable and the one with the maximum number of observations (4519) but some variables missing. Every time results differ, I will point it out.

2.3 Descriptive statistics

Table 2 and density graphics present statistical differences between boys' and girls' scores. In the remaining of the paper, all descriptive statistics and analysis are performed on standardized test scores – mean zero and variance equal one. Standardization is done within score (blind and non-blind), subject and term.

[TABLE 2 HERE]
[KERNEL GRAPHS HERE]

Graphics 1 and 2 display distributions of blind and non-blind scores during first term in French. In this subject, girls strongly outperform boys, and this premium is not affected by the nature of the grade (blind or non-blind). Girls' average score is 0.434 points higher than boys when the score is blind and 0.460 when it is non-blind. However, the story is different in mathematics. Graphics 3 and 4 show that boys outperform girls when grades are blind, but the reverse is observed when teachers assess their pupils. Hence, girls' average score during first term is 0.147 points lower than boys when the score is blind but it is 0.169 points higher when it is non-blind. Graphically, a clear shift to the right of girls' score distribution is observed (relative to boys) when comparing blind and non-blind scores in math.

Graphics 5 to 8 present girls' and boys' evolution of blind scores between the beginning and the end of school year, hence capturing their relative progress. In math, the initial boys' premium vanishes between the first and last term. Girls progress more than boys so that, by the end of the year the average gap between boys' and girls' scores in Math is no more statistically significant. One of the objectives of this paper is to determine whether part of this catching up is the result of encouragement generated by grade positive discrimination. In French, no clear difference in progress between boys and girls is observed.

3 Gender discrimination in grades

3.1 Model of grade attribution

I define a simple model to describe how blind and non-blind scores are attributed. The main assumption of this model is that blind scores are free of any bias, and hence should only measure

pupils' ability, whereas non-blind scores can be affected by teacher's attitude towards boys or girls. Hence, blind scores are modeled as a function of a pupil's ability only:

$$B_i = \theta_{1i} + \epsilon_{iB} \tag{1}$$

Here θ_{1i} is a pupil's ability, B_i is a noisy measure of a pupil's ability, and ϵ_{iB} corresponds to an individual random shock specific to blind scores. This might capture any effect that makes a pupil overperform or underperform the day of the exam and can be interpreted as measurement error. Non-blind scores can be affected by teachers' beliefs towards pupils' gender. Hence, they can be modeled as a function of both ability and pupils' gender:

$$NB_i = \alpha_0 + \theta_{2i} + \alpha_2 G_i + \epsilon_{iNB} \tag{2}$$

Here θ_{2i} is the pupil's ability that the non-blind test is meant to measure. G_i is a dummy variable that takes the value 1 for girls. α_2 is the coefficient representing the potential gender related discrimination. The constant α_0 represents the average gap for boys between the non-blind score and the ability ($NB_i - \theta_{2i}$). ϵ_{iNB} is an individual shock specific to grades attributed by teachers. This noise might capture pupils' behavior for instance. Finally, I allow θ_{1i} and θ_{2i} to differ, meaning that abilities measured by blind and non-blind scores might differ. The relationship between both abilities can be modeled as follows:

$$\theta_{2i} = \rho\theta_{1i} + v_i \tag{3}$$

Where v_i captures variables that potentially affect ability measured by class exams θ_{2i}, once controlled for ability measured by blind score θ_{1i}. Any specific ability measured by class exams but not by standardized tests, would be captured by v_i. I discuss further in the next section the importance of differentiating abilities measured by both tests. Ability measured by blind scores (θ_{1i}) might include pupils' long-term memory and their ability to synthesize knowledge acquired in the last few month, while ability measured by non-blind scores (θ_{2i}) might integrate more short-term skills such as learning an exercise by heart and replicating it the day after for the class exam. Any difference between θ_{1i} and θ_{2i} could bias the identification of discrimination. If the blind and the non-blind scores measure slightly different abilities, and if boys or girls are more endowed in one of this ability, then the coefficient α_2 of gender would not only measure a potential discrimination, but also the difference in ability distribution between boys and girls.

This way of modeling blind and non-blind scores is highly simplified and relies on two important hypotheses. Firstly, I suppose a linear relation between non-blind scores, ability and gender. Secondly, I assume that non-blind scores do not depend on blind scores in this specification. This hypothesis is likely to be satisfied in our context because blind tests were not corrected by teachers but by independent correctors.

The reduced form of this structural model is obtained by replacing θ_{2i} by its formula in equation (2):

$$NB_i = \alpha_0 + \rho\theta_{1i} + \alpha_2 G_i + (\epsilon_{iNB} + v_i) \tag{4}$$

Replacing θ_{1i} by $(B_i - \epsilon_{iB})$ gives the following reduced form:

$$NB_i = \alpha_0 + \rho B_i + \alpha_2 G_i + (\epsilon_{iNB} + v_i - \rho\epsilon_{iB}) \tag{5}$$

It is worth mentioning that this model could be used to study other sources of discrimination than gender. For instance, biases in grades linked to pupils' social background, pupils' behavior or their academic level could be studied by replacing G_i by any other interesting variable in equation (5).

3.2 Identification strategy for discrimination

To identify a potential gender bias in grades, I first use a double-differences strategy. Other papers have used this method to estimate discrimination: Falch and Naper (2013), Breda and Li (2012), Lavy (2008), Goldin and Rouse (2000) and Blank (1991). This strategy consists in estimating the difference between boys' and girls' average gap between the non-blind and the blind scores, assuming that only girls' non-blind scores suffer from stereotypes. In the absence of teachers' biases in grades, and under the assumption that both tests measure the same abilities, the difference between the non-blind score and the blind score should be the same for boys and girls. This corresponds to the common trend identification hypothesis. Implementing a double difference controls for the average effect of non-blind grading on scores, for the average effect of being a girl on score, so that what the double difference captures is the specific effect of the grade being non-blind on girls scores, relative to boys.

One of the advantages of the reduced form equation (5) is that it is compatible with an identification based on double-differences, provided that the following assumptions is made: blind and non-blind scores are assumed to measure the same abilities, so that $\theta_{2i} = \theta_{1i} = \theta_i$. In equation (5) this is equivalent to $\rho = 1$ and $v = 0$. This hypothesis is often implicitly made in other papers. I make it clear here, and will discuss its robustness in a further section, by analyzing the identification of discrimination in the more general setup where both tests do not measure the same abilities. To begin with, I consider this assumption as valid, so that equation (5) is equivalent to the usual double-differences equation:

$$NB_i - B_i = \alpha_0 + \alpha_2 G_i + (\epsilon_{iNB} - \epsilon_{iB}) \tag{6}$$

A more common formulation of this DiD specification is written below. The estimates obtained for discrimination are similar but equation (7) has the advantage to provide coefficients for the gender effect and the non-blind effect:

$$Sco_{in} = \alpha + \beta G_i + \gamma NB_i + \alpha_2 (G_i * NB_i) + \pi_c + \epsilon_{in} \tag{7}$$

Here Sco_{in} is the grade received by a pupil when the nature of scoring is n (n=1 for non-blind and 0 for blind). Hence, for each pupil, this dependent variable is a vector of both blind and non-blind grades received. G_i is a dummy variable equal to 1 if the pupil is a girl. NB_i

is a dummy variable equal to 1 if the score has been given non-anonymously by a teacher. The coefficient I am interested in is the coefficient α_2 of the interaction term which identifies gender discrimination. Finally, π_c is a class fixed-effect aimed at capturing elements affecting grades in a given class: teachers' severity for instance, or student/teacher ratio, peers effects... In further specifications, additional control variables will be added such as pupils' behavior, parents' profession, or pupils' initial level.

3.3 Empirical results on discrimination

[TABLE 3 HERE]

Table 3 presents the coefficient estimates of equation (7). Two different regressions are run in math (columns 1 and 3) and French (columns 2 and 4). In all specifications, standard errors are estimated with school level clusters to take into account common shocks at the school level. I find that in math, the coefficient of the interaction term Girl*Non-Blind is high and significant - 0.31 points of the s.d - meaning that girls benefit from a positive discrimination in this subject. This result suggests that the extent of the bias is important: girls' non-blind scores are on average 6.2% higher than boys in math during first term due to discrimination. Using the reduced sample or the whole sample does not change much the results. On the contrary, in French the coefficient of the interaction term is neither high nor significant, meaning that no gender bias is observed in this subject.

These results confirm up to a point what Lavy (2008) observes in his analysis: in opposition to what common beliefs about girls' discrimination would predict, the biases observed are in favor of girls. However, I find such a bias only in math and not in all subjects as Lavy observes. My regressions also yield larger point estimates than those found by Lavy – ranging from -0.053 in literature to -0.053 in English. My estimates also confirm those of Robinson and Lubienski (2011) who find that teachers in elementary and middle schools consistently rate females higher than males in both math and reading, even when cognitive assessments suggest that males have an advantage. The results of Breda and Ly (2012) are also consistent with my estimates. They find that discrimination goes in favor of females in more "male-connoted" subjects (e.g Math). Finally, some studies highlight that the interplay between student and teacher gender plays a role in scores biases. A recent paper by Ouazad and Page (2012) points that female' teachers give lower grades to male student, whereas male teachers tend to reward male students more than female students. Unfortunately, I have no information on teachers' characteristics; especially I do not know teachers' gender in the sample. However, I know that on average, 78% of French teachers are women while they are only 45% in math[2]. This information, together with my estimates tend to challenge Ouazad et al. results, according to which I should observe more positive discrimination toward girls in French, and a negative discrimination toward girls

[2]Source : Repères et références statistiques sur les enseignants, la formation et la recherche – 2013 - DEPP, French Ministry of Education.

in math where the majority of teachers are men.

A second coefficient is worth commenting in the DiD specification: the gender coefficient. In French, this coefficient is always positive, ranging from 0.35 to 0.42 depending on the sample used. On the contrary, in math, the coefficient is negative meaning that girls do worse than boys, conditional on the nature of scoring and discrimination. These results are interesting for two reasons: they tend to confirm stereotypes according to which girls are better in French and boys in math. These findings also contradict Lavy (2008) who finds that "overall, female high-school students had higher achievements on the state matriculation exam (blind tests) in all subject except for English". One potential justification to the difference observed is that Lavy's data relates to students in high school whereas my analysis is based on 6th grade pupils in lower secondary school. If girls catch up their initial lower level in math whilst they are in secondary school, I would no more observe the negative coefficient in high school. This catching up is observed in the second part of the analysis: the gap between boys and girls blind score vanishes between the first and last term.

3.3.1 Does gender discrimination capture girls' better behavior than boys?

A first hypothesis I want to test is whether the coefficient for gender discrimination captures girls' better behavior than boys. Controlling for pupils' punishment improves the robustness of the identification strategy compared to previous studies that had no information on pupils' sanctions. If behavior influences teachers' assessment (consciously or unconsciously), since boys are more turbulent than girls, pupils' behavior would have an unbalanced effect on non-blind scores between boys and girls. In equation (6), without any controls for pupils' punishment, the latter would enter the error term, and would be correlated with the gender variable. As far as I know, previous studies were not able to disentangle the 'pure' gender discrimination from a discrimination related to pupils' behavior. This is one of the contributions of this paper.

I create a variable "Punishment" that is a proxy for a pupil's bad behavior. It takes the value 1 if a pupil has received a disciplinary warning from the class council during first term or if he/she was temporary excluded from the school. Both punishments can be cumulated by pupils. During first term 8% of pupils received at least one sanction: 6.2% received a disciplinary warning and 3.6% were temporary excluded from the school. Boys are more punished than girls: among pupils having at least one sanction during the first term, 84.4% are boys and 15.6% are girls. Several schools did not provide information on their pupils' behavior, so that for many pupils the punishment variable is missing. Following regressions will focus on the sample of 2269 pupils for which punishments are non-missing[3]. This sample being different from the previous one, I run a balance check to verify if pupils' characteristics differ. I find no significant difference regarding the blind score, non-blind score, gender and parent's profession. Even if schools which do not provide information on sanctions are the one with the most turbulent students, my results

[3]The sample is the full sample, to which pupils for which a punishment variable is missing have been removed.

will be a lower bound of the effect of pupils' behavior on the gender bias.

[TABLE 4 HERE]

Results are presented in table 4, column 2. Regressions are run in math, where gender discrimination is observed, with the sample of pupils for which the punishment variable is not missing. To ensure that coefficient comparisons are based on the same sample, column 1 presents results of the standard DiD regression implemented on the new sample. The coefficient for discrimination decreases when I control for pupils' behavior, but the drop is small: the point estimate goes from 0.325 to 0.316 points of the s.d. This suggests that in math, around 4% of the gender discrimination I observe is due to girls' better behavior than boys.

Finally, independently from any gender bias, the coefficient of the interaction term Punishment*Non-Blind is of particular interest. It can be seen as a discrimination related to bad behavior, in the same way as I interpret the coefficient Girl*Non-Blind as a gender discrimination. My estimates suggest that a bad behavior has no effect on teachers' assessment. Once the effect of gender, non-blind grading, and teachers' gender biases are controlled for, pupils that have been sanctioned at least once during the term, have a non-blind score that is 0.14 points lower than the non-blind score of pupils who behave well, but this coefficient is not statistically significant. Teachers do not seem to use grades as messages about their like or dislike of the student's attitude.

I take into account pupils' bad behavior by including a variable that controls for punishments but girls' behavior might also affect non-blind scores through more diffuse aspects of their behavior: how they behave in the classroom, how often they answer questions, the diligence they show in their work. I consider that these elements will not bias the results as long as they are a component of my definition of girls. In this case, the coefficient for gender discrimination captures some characteristics that are intrinsically linked to girls.

3.3.2 Does gender discrimination in math capture girls' initial lower achievement than boys?

The second hypothesis I want to test is whether discrimination towards girls partially captures two potentially related effects: (1) some teachers' might give more favorable grades to low-achievers and (2) in some classes the variance of the teachers' grades might be smaller than the variance of the standardized scores. Firstly, some teachers might behave differently towards low-performers, and potentially give them higher grades than expected by their ability. We have seen in descriptive statistics that girls outperform boys in French but they perform lower than boys in math. If teachers have a tendency to give good grades to low-performers to encourage them, then what I interpret as gender discrimination could partially capture a 'low-achiever' positive discrimination. Secondly, some teachers might have a lower dispersion of their attributed grades than the dispersion of the standardized scores. For a given dispersion of blind scores in a classroom, reducing the dispersion of non-blind scores will improve the non-blind score of the weakest in the class relatively to the scores of the best pupils. Again, since girls have initially

lower scores than boys in math, a teacher who prefers a reduced dispersion of his grades will advantage girls compared to boys.

To test these hypotheses, I first add controls for pupils' initial position in the blind grade distribution. The new specification includes dummy variables indicating whether pupils belong to the lowest or highest 10th deciles. I decompose blind scores into deciles rather than non-blind scores. The former seem more relevant than grades attributed by teachers because they are independent from teachers' assessment - hence evaluate pupils' ability more trustworthy. Scores are decomposed into deciles within each subject and within class, meaning that pupils are ranked relatively to other children in their class.

The column 4 in table 4 presents results when a variable controlling for low achievers is included (pupils below the 1st decile) and column 5 presents results with a variable controlling for high achievers (pupils above the 9th decile). The point estimate of gender bias decreases by 7.5% when controls for low achievers are added to the regression – from 0.318 to 0.294 - suggesting that part of the gender bias in math captures a low-achiever bias. As a second test, I run the regression on pupils' rank instead of pupils' test scores. Teachers' narrower or larger dispersion of their grades does not affect their pupils' ranking within the class. Hence running DiD regressions with pupils' rank as a dependant variable is a mean to control for teachers' preferences. Table 9 displays the coefficients of these regressions, which I run on the initial whole sample containing 8329 observations in math and 8315 in French. Coefficients are consistent with previous conclusions: the coefficient of the interaction term equals -2.2, meaning that girls' average rank decrease by 2.2 when they are assessed by their teacher – going from 22 to 19.8 for instance.

[TABLE 5 HERE]

Finally, the results in table 4 suggest that, independently from any gender discrimination, a positive bias exists in favor of low-achievers. In math once I control for gender, non-blind grading and the gender bias, belonging to the 1st decile has a positive impact on non-blind scores (compared to blind scores) of 0.261 points of the s.d. This result is relevant with an encouragement effect from teachers in favor of low-achievers. The magnitude of the effect implies that pupils belonging to the lowest decile receive grades 5.13% higher than the other pupils, due to the positive discrimination they benefit from. For high achievers, I observe a reversed situation. Belonging to the the 10th decile has a negative effect on non-blind scores relative to blind scores, once controlled for gender and non-blind grading. In order to know how the magnitude of this bias differs across deciles, I run the previous regression by adding an interaction between the non-blind score and each decile – the 6th decile being taken as a reference. The graphic 9 plots the distribution of point estimates of the interaction term between the non-blind score and each decile. It suggests that the positive bias in favor of low-achievers is all the most important as they belong to the lowest decile, and reversely that the negative bias towards high-achievers is increasing as pupils belong to the highest deciles. Similar results are found in French.

[GRAPHIC 9 HERE]

Two interpretations seem plausible. Firstly, as discussed above, teachers might have a preference for homogenous grades. They might prefer a low dispersion of grades among their pupils, and hence tend to overgrade pupils with more difficulties and to be more severe with good pupils. Secondly, these coefficients could be interpreted as a mean reversion effect in so far as the blind and the non-blind scores are not measured at the same date – the blind test is taken before the non-blind. Low achievers may respond more vigorously than high achievers to the grade they receive and make a relatively greater improvement as a consequence. However, this interpretation is unlikely because pupils were not informed about the score they obtained for the standardized blind test[4].

3.4 Robustness checks

3.4.1 Are both tests measuring the same abilities?

The DiD specification discussed above rests on the restrictive assumption that both tests measure the same abilities. However, if blind and non-blind scores do not measure exactly the same abilities, and if these skills are not equally distributed between boys and girls, then failing to take it into account will yield biased DiD estimates of gender discrimination. In equation (6), the coefficient α_2 which I interpret as discrimination would partly capture girls or boys specific ability in blind or non-blind scores. In this paper, I am extremely careful about this concern since blind tests are standardized tests created by the French Education Ministry, while non-blind grades correspond to the average mark given every term by the teacher. They might measure slightly different abilities. In mathematical terms the assumption that both tests measure the same ability is equivalent to $\rho = 1$ and $v_i = 0$ in equation (3) defined previsously: $\theta_{2i} = \rho\theta_{1i} + v_i$. If we release this hypothesis, we are back to the reduced form equation presented previously:

$$NB_i = \alpha_0 + \rho B_i + \alpha_2 G_i + (\epsilon_{iNB} + v_i - \rho\epsilon_{iB})$$

A way to test the validity of the hypothesis is to directly estimate the reduced form equation above and to verify if the coefficient ρ is significantly different from one. If not, both tests can be assumed to measure abilities which are perfectly correlated and DiD estimates can safely be assumed to be unbiased[5]. However to correctly estimate the parameter ρ in this equation, I have to get rid of the measurement error bias on B_i. Since B_i is a noisy measure of ability θ_{1i}, it is correlated to the measurement error ϵ_{iB}. I solve this endogeneity issue by instrumenting B_i. A pupil's month of birth is used as an instrument that is correlated to his/her blind score but independent from the error term.

[4] Finally, I test whether parents' profession has an impact on discrimination and find no significant effect of pupils' social background.

[5] I will discuss in a further section an additional assumption required for the DiD to be unbiased. Although we cannot test whether $v_i = 0$, the term v_i should be equally distributed between boys and girls.

In the literature, students' month of birth has been shown to be an important determinant of pupils' success at school (Crawford et al. 2007, Bedard and Dhuey 2006 and Grenet 2012). I test the correlation between blind scores and pupils' month of birth by running a regression of blind scores in French and math on a set of 11 dummies for each month of birth. January is taken as the reference month so that all coefficients should be interpreted relatively to this month. The graphic 10 presents the correlation coefficients.

[GRAPHIC 10 HERE]

There is clear evidence that pupils born at the end of the year have lower results than those born between April and September. These coefficients suggest that the month penalty begins in August for scores in French, and September in math. From this observation, and to avoid including too many instrumental variables in the equation, I compute a new instrumental variable that is a dummy for pupils born at the end of the year in each subject (between September and December in math for instance). Results of the first stage regression are displayed in table 6. Once controlled for covariates, being born at the end of the year has an important negative effect on blind scores – 0,166 points of the s.d in math and 0,180 in French.

[TABLE 6 HERE]

Being born at the end of the year will be a valid instruments if the following exclusion restriction holds: the only reason why a pupil's month of birth affects teachers' grades is because being born at the end of the year impacts his ability – measured by the blind score – once controlled for other covariates. In other words, being born at the end of the year is uncorrelated to the random shocks that enter the error term of equation (5):

$$Cov(BornEndYear_i; \epsilon_{iNB} + v_i - \rho\epsilon_{iB}) = 0$$

I claim this restriction is valid, provided that I control for pupils' behavior, parents' profession and grades retention, three variables that might be correlated to being born at the end of the year. The reduced form equation (5) is estimated, first with standard OLS, and second by instrumenting the blind score. Results are presented in table 7. As previously, all regressions include class fixed-effects. They are run on a sample that contains 2168 pupils in math for which the blind score, non-blind score and punishment variable are non-missing, and 2120 in French. Standard errors are estimated with school level clusters to take into account common shocks at the school level.

[TABLE 7 HERE]

Regarding the coefficient of interest α_2, I observe that the estimation with IV – 0.338 in math and 0.064 in French – is very similar to the coefficient I obtain by implementing DiD – 0.323 and 0.043 respectively. This confirms my results suggesting a bias in teachers' grades in favor of girls.

Second, the purpose of this estimation is to check whether both tests measure abilities which are perfectly correlated, in other words if the IV coefficient of the blind score is equal to one. This coefficient ranges from 1.015 in French to 1.082 in math and in both cases I cannot reject the hypothesis that $\rho = 1$. This result suggests that the blind and non-blind tests measure skills that are perfectly correlated, and hence that implementing double-differences gives unbiased estimates of gender bias. Hence, for further robustness checks, the DiD specification will be used.

Some might argue, regarding the exclusion restriction, that once controlled for the abilities measured by the blind score, being born at the end of the year is not perfectly independent from unobserved specific skills v_i tested by the non-blind score only. If this is the case, it is likely that being born at the end of the year would also be negatively correlated with these unobserved skills. Therefore the IV estimates of ρ might be an upper bond for the true value of ρ[6], while the OLS would be a lower bound (due to the downward measurement error bias). What is most important to me is how the potential bias on the coefficient ρ affects my estimate of interest α_2 in the reduced form presented above: $NB_i = \alpha_0 + \rho B_i + \alpha_2 G_i + (\epsilon_{iNB} + v_i - \rho\epsilon_{iB})$.
Since girls perfom initially lower than boys in mathematic, and higher than them in French, the blind score is correlated to a pupils' gender, so that the bias on ρ could affect the estimate of α_2. Using the omited variable bias formula, we can easily show that the downward bias on ρ creates a downward bias on α_2 in maths, but an upward bias in French. The direction of this bias is fully determined by the correlation between pupils' sex and the blind score - as detailed in appendix 1. This downward bias on α_2 implies that, even if my estimates using DiD and IV might overestimate the real gender bias, the lower bound of this effect is still high and significant: the OLS estimate of α_2 equals 0.262 in math and 0.173 in French. This confirms that in math a substantial bias exists in favor of girls. In French, the coefficient should be interpreted more carefully. The upper bound of the gender bias finds a significant effect, but any other method aimed at reducing the bias do not find any significant gender bias.

3.4.2 Could girls progress more than boys between the date of the blind test and the date of the non-blind?

Pupils take the standardized blind test during one of the first days of the school year whereas teachers' assessment is an average of several grades given by teachers during the first term. Since the first term lasts three month, this average of several grades measures a pupils' average ability about one and a half month after the beginning of the school year. This time lag between the date of the blind and non-blind scores might be problematic if girls tend to progress more than boys during this period. In particular, if teachers' biases in math appear early in the school year, it might affect girls' progress from the first weeks of the school year. In this case, the coefficient

[6] The IV estimate is $\rho_{IV} = \frac{Cov(NB_i, EndYear_i)}{Cov(B_i, EndYear_i)}$. If a correlation exists between v_i and being born at the end of the year, this would affect the numerator of the formula by increasing $Cov(NB_i, EndYear_i)$. Hence ρ_{IV} would be an upper bound for the parameter ρ.

which I interpret as a gender bias in math would be an upper bound for the true gender bias.

To address this concern, I use the data that have been collected at the end of the academic year. Fortunately, the same scores have been collected - standardized tests and teachers' given grades – but the time lag is reversed during the last term. Pupils take the standardized blind test during one of the last days of the school year, while teachers' assessment is an average of several grades given by teachers during the three last months. Hence, the blind test is taken after the non-blind test. Under the same assumption that girls tend to progress more than boys during this period, my estimates of gender discrimination during third term would be a lower bound. Computing the lower and upper bound of the estimates enables us to find a plausible interval for the gender bias.

[TABLE 8 HERE]

I run the same DiD regression as before but on the third term scores. Then I compare the estimates obtained during first term (upper-bound) and last term (lower bound). Results are displayed in table 8. The same full sample is used for both regressions. Consistent with the hypothesis that girls progress more than boys in math, the third term coefficient (0.251) is lower than the first term coefficient (0.318). The true value of gender discrimination is likely to be between 0.251 and 0.318.

3.4.3 Could girls be more affected by some unobserved shocks ?

The simple model defined in section 3 contains three unobserved shocks: (1) ϵ_{iB} corresponds to an individual shock specific to blind scores, (2) ϵ_{iNB} corresponds to an individual shock specific to non-blind scores and (3) v_i captures any specific ability measured by class exams but not by standardized tests. The DiD estimates rest on the assumption that these shocks are equally distributed for boys and girls[7]. However, if girls are systematically more stressed than boys for standardized tests, if they tend to attach more importance to these tests, or if they are more endowed in specific abilities measured by class exams, the restrictive assumption would be violated and the DiD estimates could be biased.

Firstly, one might believe that girls are more stressed than boys for standardized tests, so that they tend to underperform in this kind of examination. My coefficient of discrimination would be an upper bound for true gender discrimination. However, two reasons make me believe that girls were not more stressed than boys in this context. Firstly, both tests are taken in the same conditions: pupils take the standardized test and their class exam in the same classroom where they seat usually, and it is their teacher who gives the instructions. What is more, standardized tests are not high-stakes for the students. A pupil's result to this test is not accounted for to compute his/her end of term average score. Secondly, it might be that girls are more endowed in abilities which are measured by class exams but not by standardized tests.

[7] In mathematical terms, this means that $E(\epsilon_{iNB}/G_i = 1) = E(\epsilon_{iNB}/G_i = 0), E(\epsilon_{iB}/G_i = 1) = E(\epsilon_{iB}/G_i = 0)$ and $E(v_i/G_i = 1) = E(v_i/G_i = 0)$.

These abilities could recover short-term memory or learning an exercise by heart and replicating it the day after for the class exam. For both examples – stress and abilities - if girls were systematically underperforming at standardized tests because of the shock and if this shock was equally distributed between subjects, then girls should also have a lower blind than non-blind score in French. I do not observe this. In French, the gap between the blind and non-blind score for girls is the same as the one for boys.

Comparing the coefficient for discrimination in math and French, as I do here, is equivalent to implementing within-gender between-subjects regressions – or triple differences (Breda et al, 2013). This is a mean to control for any unobserved shock or characteristics that differ across gender but are assumed to be constant between subjects. Typically, triple differences allow v_i to be distributed differently for boys and girls, but within gender v_i must be constant between French and math[8]. The coefficient for relative discrimination obtained with this method corresponds to the coefficient in math minus the one in French, hence 0.291 for the whole sample. I still conclude that a positive bias exists in math in favor of girls. Finally, it should be noted that this within-pupil between-subjects method controls for any characteristic specific to girls that potentially affect teachers' biases: the fact that girls are less turbulent, might be more attentive, more serious, more diligent...

4 Impact of discrimination on pupils' progress

My results on gender discrimination pave the way for a new set of questions related to the impact of discrimination on pupils' motivation and on their achievement at school. Positively discriminating students might encourage them to do more efforts, and hence to increase their scores. Reversely, if achievement and efforts are substitutes, some students benefiting from positive discrimination could provide less effort considering that they are good enough (Benabou and Tirole, 2002). The dataset I use has the benefit to contain data on scores at two different periods in time - at the beginning and at the end of the 6th grade. This allows me to analyze the effect of gender discrimination on students' progress.

4.1 Comparisons of girls and boys progress

[GRAPHICS 11 AND 12 HERE]
[TABLE 9 HERE]

Graphics 11 and 12 plot the distribution of boys and girls progress between the first and the last term. I define progress as the difference between the blind score at the end of school year and the blind score at the beginning of the year. This difference can be interpreted as a pupil's progress because both standardized tests measure the same abilities. They are designed by the

[8] In mathematical terms, this means that $E(v_{i,french}/G_i = 1) = E(v_{i,math}/G_i = 1)$ and $E(v_{i,french}/G_i = 0) = E(v_{i,math}/G_i = 0)$.

French Ministry of Education and aimed at measuring the same abilities at two different periods in time.

Graphically, there is clear evidence that girls progress more than boys in mathematics, whereas progress in French is similar. In math during first term, girls' average score was 0.069 points below the mean, while it is 0.012 points above the mean during last term, hence an increase of 0.082 points of the s.d. Since girls' blind scores were lower than boys' at the beginning of 6th grade, the fastest progress experienced by girls reduces the gap between boys and girls blind scores. Girls are catching up boys in math. These different patterns of progress between boys and girls in math raise the question of the link between the positive bias in grades I observe towards girls in this subject and their subsequent higher progress.

4.2 Model of pupil's progress

I define a simple model aimed at isolating the effect of teachers' biased assessment on pupils' progress. To begin with, I will keep the model as general as possible so that discrimination could be considered towards any group of pupils. The main issue when evaluating the impact of grade discrimination on a pupil's progress is to disentangle the pure effect of grade biases from several other determinants that might explain a pupil's high or low progress: how much of the progress is due to discrimination? How much is due to specific characteristics of the discriminated group? For instance, girls might have an intrinsic tendency to progress more than boys over the school year, without any discrimination. Similarly, low-achievers might have an initial higher propensity to progress than high-achievers, again independently from any discrimination. Finally, I want to take into account the fact that some teachers are more able than others to make their entire class progress. The following model aims at isolating these various determinants of pupils' blind scores evolution over the school year. Equation (8) below describes blind scores during first term (as defined in section 2.1), while equation (9) describes blind scores during last term:

$$B_{1i} = \theta_{1i} + \epsilon_{B1i} \tag{8}$$

$$B_{3i} = \theta_{3i} + \epsilon_{B3i} \tag{9}$$

For the remaining of the model, all variables and parameters for third term are indexed by 3. A pupil's ability has changed between the first and the last term. I model third term ability as a function of the three effects I want to disentangle: the effect of discrimination, a pupil's independent tendency to progress compared to the others and a teachers' effect on progress:

$$\theta_{3i} = \delta\theta_{1i} + \alpha G_i + \mu_i T_i + \beta D_{1i} + \omega_i \tag{10}$$

Third term ability θ_{3i} depends on three potential effects: (1) a discrimination effect caused by teachers' biased assessment of their pupils: βD_{1i}, where D_{1i} corresponds to grade discrimination during first term. Its impact on pupils' third term ability is measured by the coefficient β. It

is important to understand that this coefficient captures several channels through which grade biases can affect a pupil's third term score. Motivation or discouragement are direct channels, but effort is also an important channel, as well as change in self-confidence and reduction of stereotypes threats. I will not be able to distinguish between these different channels, that are all captured by the coefficient β. (2) Second, third term ability θ_{3i} also depends on the independent tendency to progress of the discriminated group, relatively to other pupils. This is captured by the coefficient α. In this general model, G_i is a dummy variable that equals one for pupils belonging to the discriminated group. In a model where only gender discrimination is considered, G_i would correspond to a girl dummy. (3) Finally, a pupil's progress is affected by his/her teacher ability to make the entire class progress, where T_i is a teacher dummy.

Compared to the model of discrimination presented in section 3, I assume here that the blind and non-blind tests measure the same abilities during first term. This assumption is based on results obtained in the first part. Following the first robustness check, I could not reject the hypothesis that both scores are measuring skills that are perfectly correlated.

In equation (10), I replace the coefficient for discrimination D_{1i} by $NB_{1i} - \theta_{1i}$, which corresponds to the difference between a pupil's ability and the non-blind grade attributed by his/her teacher during the first term. This corresponds to discrimination during first term. Equation (10) becomes:

$$\theta_{3i} = \delta\theta_{1i} + \alpha G_i + \mu_i T_i + \beta(NB_{1i} - \theta_{1i}) + \omega_i \tag{11}$$

By replacing θ_{3i} by its expression in equation (11) I obtain:

$$B_{3i} = \delta\theta_{1i} + \alpha G_i + \mu_i T_i + \beta(NB_{1i} - \theta_{1i}) + \omega_i + \epsilon_{B3i} \tag{12}$$

Finally, replacing θ_{1i} by its equation gives the following reduced form of the model:

$$B_{3i} = (\delta - \beta)B_{1i} + \beta NB_{1i} + \alpha G_i + \mu_i T_i + [\omega_i + \epsilon_{B3i} + (\beta - \delta)\epsilon_{B1i}] \tag{13}$$

This reduced form equation isolates the effect of discrimination β, the discriminated group's independent tendency to progress α, and μ_i the teacher's effect. By rewriting it as below, the interpretation of the coefficients becomes straightforward: once controlled for a pupil's ability B_{1i}, for a group tendency to progress G_i, and for a teacher's average effect T_i, the coefficient β of the difference between non-blind and blind scores captures the effect induced by the fact that a pupil receives a grade higher than expected by his/her ability:

$$B_{3i} = \delta B_{1i} + \beta(NB_{1i} - B_{1i}) + \alpha G_i + \mu_i T_i + (\omega_i + \epsilon_{B3i} + (\beta - \delta)\epsilon_{B1i}) \tag{14}$$

4.3 Identification of girls' relative progress due to grade biases

The model defined above is compatible with any kind of grade discrimination (related to gender, ethnicity, achievement, behavior...). To build upon the results found in part 1, I will focus

now on the identification of girls progress (relative to boys) related to gender discrimination only. Therefore, in equation (14), the group dummy G_i becomes a dummy for girls. The term discrimination will always refer to gender biases in the remaining of this section. The identification strategy is based on the observation that not all teachers discriminate, and that among teachers who have a biased assessment of girls compared to boys, the degree of the bias also differs across teachers, with some teachers discriminating more than others. I take advantage of this heterogeneity in the degree of discrimination to implement a between-class analysis. It is the variance in teachers' discriminatory behavior that will identify the causal effect of teachers' biased assessment on pupils' achievement. I want to see if classes in which girls benefit from a high degree of discrimination (relatively to boys) are also classes in which girls progress more (relatively to boys). This identification strategy can be seen as a DiD strategy were the treatment corresponds to discrimination towards girls in some classes and the outcome is girls average third term blind score compared to boys.

It is worth mentioning that the impact of gender discrimination I estimate with this specification captures different elements. Teachers that tend to favor girls in their grades are also likely to have a behavior towards girls that differs from teachers who do not have biased grading. Typically, they might be more encouraging, friendlier, focus more attention on girls, or be less critical. The effect of gender discrimination on progress will capture all these effects. Even without being able to separately identify these elements, it is interesting to know if teachers' biased behaviors – with all elements it embeds – have an impact on girls' progress relative to boys.

Graphics 13 and 14 provide a good insight into this question. For each class in the sample, these graphs display girls' relative discrimination and girls' relative progress (both with respect to boys). The former is measured as the class average difference between the non-blind and the blind scores for girls, minus this same difference for boys. It corresponds to the estimate of gender discrimination obtained with the DiD in part 1. Girls' progress relative to boys is measured as the difference between their blind score at the end of the year and this blind score at the beginning of the year, minus this same difference for boys. Graphically, there is clear evidence of a positive correlation between the degree of discrimination and the degree of progress, and this is true in both French and math. It is also interesting to see that in part 1, the results suggest that on average there is no discrimination in French. Graphic 13 clearly shows that despite this null average, there is an important variance in teachers' biased assessments, which might yield girls' higher or lower progress in these classes.

[GRAPHICS 13 AND 14 HERE]

The identification strategy is based on the comparison of mean scores between classes. Based on equation 14, this requires to aggregate scores at the class level for both girls and boys. Within a class, girls' average third term blind score is given by:

$$E(B_{3i}/T_i, G_i = 1) = \delta E(B_{1i}/T_i, G_i = 1) + \beta E(NB_{1i} - B_{1i}/T_i, G_i = 1) + \alpha E(G_i/T_i, G_i = 1)$$
$$+ \mu_i E(T_i/T_i, G_i = 1) + E(\omega_i/T_i, G_i = 1) + E(\epsilon_{B3i}/T_i, G_i = 1) + (\beta - \delta)E(\epsilon_{B1i}/T_i, G_i = 1) \quad (15)$$

All variables are averaged conditionally to being a girl and having teacher T_i. Symmetrically, boys' average score within a class is given by:

$$E(B_{3i}/T_i, G_i = 0) = \delta E(B_{1i}/T_i, G_i = 0) + \beta E(NB_{1i} - B_{1i}/T_i, G_i = 0) + \alpha E(G_i/T_i, G_i = 0)$$
$$+ \mu_i E(T_i/T_i, G_i = 0) + E(\omega_i/T_i, G_i = 0) + E(\epsilon_{B3i}/T_i, G_i = 0) + (\beta - \delta)E(\epsilon_{B1i}/T_i, G_i = 0) \quad (16)$$

From these two equations, I obtain the difference in progress between boys and girls in class c[9]:

$$(B_{3G} - B_{3B})_c = \alpha + \delta(B_{1G} - B_{1B})_c + \beta[(NB_{1G} - B_{1G}) - (NB_{1B} - B_{1B})]_c + (\omega_G - \omega_B)_c \quad (17)$$

Equation 17 corresponds to the equation aggregated at the class level which I want to estimate to identify the effect of gender discrimination on progress. It is specified as a differentiation between boys and girls average scores at the class level, so that teachers' effects disappear; they affect similarly boys and girls within a class. The double difference at the right hand side of the equation corresponds to the coefficients for gender discrimination estimated in section 3 of the paper – although here there is one coefficient per class. It is also worth noticing that in equation 17, assuming $\delta = 1$ transforms it into a standard DiD equation:

$$(B_{3G} - B_{3B})_c - (B_{1G} - B_{1B})_c = \alpha + \beta[(NB_{1G} - B_{1G}) - (NB_{1B} - B_{1B})]_c + (\omega_G - \omega_B)_c \quad (18)$$

The coefficient β obtained with this DiD specification corresponds to the slops of regressions lines displayed in graphics 13 and 14. For the remaining of the analysis, I prefer keeping equation 17 which requires less restrictive assumptions. In this equation, the coefficient β identifies the effect of being assigned a teacher who discriminates girls more or less – relatively to boys – on girls' average third term blind score – relative to boys – once I control for the initial average difference between boys and girls' blind scores. This coefficient can be seen as a causal effect under the assumption that girls' assignment to a teacher who discriminates is quasi-random. In other words, being assigned a teacher who discriminates is independent from girls' unobserved characteristics ω_i that make them potentially progress more than boys, once their initial level is controlled for. I use the term quasi-random to describe the fact that pupils' assignment to

[9] Where to simplify notations:

$$B_{3G} = E(B_{3i}/T_i, G_i = 1), B_{3B} = E(B_{3i}/T_i, G_i = 0)...$$
$$\omega_G = E(\omega_i/T_i, G_i = 1) + E(\epsilon_{B3i}/T_i, G_i = 1) + (\beta - \delta)E(\epsilon_{B1i}/T_i, G_i = 1)$$
$$\omega_B = E(\omega_i/T_i, G_i = 0) + E(\epsilon_{B3i}/T_i, G_i = 0) + (\beta - \delta)E(\epsilon_{B1i}/T_i, G_i = 0)$$

teachers is not done through a proper lottery. Yet, an arbitrary assignment of girls with high predicted progress to teachers who discriminate is highly plausible for several reasons. Firstly, pupils considered in this study are in 6th grade, which corresponds to the first year of lower secondary school. When deciding the composition of classes, school heads and teachers have very little information on these new pupils, in particular it is very unlikely that they can predict their progress, and therefore influence their assigned class and teacher. Secondly, assigning teachers who discriminate to girls who have a high probability to progress more than boys would necessitate that school heads know who are the teachers who discriminate girls, which is again unlikely.

Although it is not possible to test this independence assumption, I test if the assignment to a teacher who discriminates is independent from boys and girls observed characteristics. To do so, I first regress the discrimination coefficient (defined at the class level in both French and math) on pupils' gender and find no significant effect: girls are not more assigned to teachers with a high bias than boys. This is true in French and math. Then, for both boys and girls separately, I successively regress the discrimination coefficient on the following set of variables : having upper class parents, having lower class parents, having repeated a grade. I find that these observed characteristics are independant from being assigned a teachers with a high level of bias. The only exception is that boys with upper class parents are slightly less likely to be assigned a teacher who discriminates in math, and that girls having repeted a grade are less likely to be assigned a teacher who discriminates in French. Finally, I argue that being assigned a teacher who discriminates is independent from girls' and boys' averaged random shocks affecting blind scores during first and last term. As long as these shocks recover pure testing noise – being ill the day of the exam for instance - it is plausible that they are independent from teachers' assignment.

The identification I use is based on the heterogeneity in teachers' discriminatory behaviors between different classes. It is equivalent to implement an IV strategy based on equation 13, where the term $(NB_{1i} - B_{1i})$ would be instrumented by all the interactions between teachers and girls at the class level. These interactions measure teachers' biased grading in favor of girls. The assumption detailed above - pupils' assignment to a teacher who discriminates is random - is analogous to an exclusion restriction on these instrumental variables.

Before turning to the empirical results, it is worth mentioning three advantages of this between-class comparison, relatively to an estimation of parameters with individual observations based on equation 13. Firstly, comparing classes rules out the issue of girls' potential higher stress than boys for blind tests. Here the double-differences nature of equation 17 implies that any effect that is common to all classes disappears. As long as pupils' assignment to teachers who discriminate is independent from their unobserved characteristics that make them progress more, then girls with higher stress for standardized tests should be equally distributed between classes. A second concern when analyzing discrimination and progress with individual observation is the potential for reversed causality caused by the fact that teachers might discriminate more

pupils they believe have an ex-ante high potential for progress. In my setting, the arbitrary assignment of pupils implies that those with an ex-ante high potential for progress should be equally distributed between classes. Hence, comparing classes rules out this problem. Finally, averaging scores at the class level reduces significantly the measurement error affecting blind score when measured at the individual level.

4.4 Empirical results on girls' relative progress

The first regression is based on equation 15. Six scores are included in this regression, so that our sample is automatically restricted to classes for which none of these scores is missing. In math the sample contains 175 classes out of 191, 171 in French. The classes excluded from the analysis do not differ significantly regarding the gap between boys and girls' blind score during first term, or in term of teachers' discriminatory behavior. Results are displayed in table 10.

[TABLE 10 HERE]

The key result suggests that classes in which teachers present a high degree of discrimination in favor of girls, are also classes in which girls tend to progress more over the school year compared to boys. The coefficient is high (0.287) and significant. In a class where boys and girls would have on average the same initial blind score, positively rewarding girls by increasing their non-blind score by one point of the s.d compared to boys, would increase the gap between boys and girls third term blind score by 0.28 points of s.d. I can also interpret this coefficient in light of my first part results. I found an average discrimination coefficient of 0.31 in math, which implies that, proportionally, girls' third term blind score would be 0.089 points - or 1.9% - higher than boys.

To build upon these results, it is interesting to see whether the catching up of girls we observe in math would still have happened in the absence of gender discrimination. The descriptive statistics presented in table 2 show that, during third term in math the gap between girls and boys blind score equals -0.040 points of the s.d, while it is -0.147 during first term. This represents a relative improvement of girls compared to boys of 0.107 points of the s.d. My results suggest that, in the absence of gender discrimination, girls and boys gap during third term would have been equal to -0.129 instead of -0.040, therefore a relative improvement of girls of 0.018 instead of 0.107. This suggests that, in the absence of discrimination, girls would not have progressed more than boys, so that the catching up we observe in math is almost entirely driven by the positive effect of grade discrimination on girls' progress. The effect of grade biases on progress is observed in math but no significant effect is observed in French. This result is consistent with graphic 10 which shows that in French girls do not progress more than boys over the school year.

We should be careful when interpreting the coefficient and keep in mind that the outcome is relative. It corresponds to the difference between girls and boys scores, so that the positive coefficient I find could correspond to a higher progress for girls than for boys, but it could also correspond to a blind score that remains constant for girls between first and last term, but

decreases for boys due for instance to their feeling of being negatively discriminated compared to girls.

These findings are consistent with different mechanisms mentioned in prior literature. Firstly, positively rewarding girls can reduce the stereotype threat effect. In situations where stereotypes are perceived as important, some girls have been proved to perform poorly for the sole reason that they fear confirming the stereotypes (Spencer et al. 1999). If math is perceived by girls as more affected by teachers' stereotypes, over-grading girls can reduce their anxiety to be judged as poor performers, and therefore favor their progress. The fact that biases affect girls' relative progress in math but not in French is consistent with a reduction of the stereotype threat, which might be more prevalent in math than in French. My findings are also consistent with prior research highlighting a 'contrast effect' according to which a student's academic self-concept is positively influenced by his or her individual achievement, but negatively affected by other peers-average achievement - usually composed of peers in the classrooms - once controlled for individual achievement (Trautwein et al. 2006, Marsh and Craven, 1997). With regard to this contrast effect, giving higher grades to girls would have a twofold effect: from an absolute point of view, higher grades will positively affect girls' self-concept, and self-confidence in math, and from a relative point of view, girls' higher grades compared to boys will reduce the achievement gap between boys and girls, and therefore increase girls relative academic self-concept.

Finally, my result tends to challenge Mechtenberg's (2009) theoretical predictions according to which girls are reluctant to internalize good grades in math, because they believe their grades are biased. Girls' mistrust of their grades may disincentive them to provide more efforts to improve their scores. This would explain why boys outperform girls in math. According to Mechtenberg, the lowly talented girls partially internalize the reward in math, and therefore achieve more on average than the lowly talented boys. But this effect is compensated by the highly talented girls who "mistrust the praise they get and achieve therefore less on average than the highly talented boys". A key feature in Mechtenberg's model is students' beliefs about their teachers' grading practices – whether they are biased or not. This might be one reason why my results do not coincide with Mechtenberg's predictions. The pupils in this study are in 6th grade, which is the first grade of lower secondary school. The year before, all pupils were attending a different school so that, when they arrive in their new school, they have no information on teachers, in particular it is very unlikely that they know which teacher discriminates.

5 Conclusion

In most OECD countries, boys outperform girls in mathematics, but they underperform in humanities. Then over the school years, this achievement gap vanishes in math but persist in French. This paper studies a potential explanation of this persistent achievement gap: teachers' biased grading practices and show that the effect of teachers' biases on girls' progress is consistent with the reduction of the gender gap observed empirically in math. I use data containing both

blind and non-blind scores at different periods in time to identify first the effect of teachers' stereotypes on their grades, and second the effect of biased rewards on pupils' progress over the school year. Firstly regarding discrimination, my results suggest that an important positive discrimination exists in math towards girls, while no bias is observed in French. A small fraction of this gender bias captures girls' better behavior than boys, and girls initial lower achievement in math. Regarding the impact of discrimination on girls' progress relative to boys, I observe that classes in which teachers present a high degree of discrimination towards girls, are also classes in which girls tend to progress more over the school year compared to boys, hence suggesting a positive effect of rewards on girls' relative progress in math. Thus, my results provide new empirical evidence on gender discrimination in grades and how it affects the gender achievement gap.

I am however unable to disentangle the different channels through which a gender bias can affect girls' relative achievement. On the one hand, positively rewarding pupils could motivate them, make them increase their efforts, increase their self-confidence, and reduce the stereotype-threat they suffer from. On the other hand, if pupils consider effort and abilities as substitutes, a higher grade might be an incentive to reduce effort and work. Unfortunately, I am not able to disentangle these effects that might compensate or reinforce each other. This is an interesting question for future research. Another concern is the validity of my results. In this study, I use a dataset that has been collected in schools of a relatively deprived educational district. This must be considered for issues of external validity of this analysis. Teachers assigned to deprived areas are on average younger than teachers in more advantaged schools. Some might argue that young teachers are less aware of stereotypes that might bias their grades, and hence be more affected by it. Having no information on teachers, I am not able to verify this point but prior research seems to conclude that the impact of teachers' experience on their grading practices is unclear and differs depending on the subject considered (Lavy, 2008). In math, the bias is found to originate mainly from older teachers.

Finally, this analysis provides several policy-relevant results regarding teachers grading. Firstly, my findings suggest that marks given by teachers do not reflect only pupils' ability. They are affected by pupils' characteristics or attitudes. This raises the question of the relevance of some elements included in grades. Should a grade reflect a pupil's gender, his/her initial achievement, or behavior? The answer is not clear and seems to depend on the objective pursued. On the one hand, if grades are wished to measure only a pupil's ability, then the influence of gender or behavior might be problematic. On the other hand, if grades are considered as a mean to counterbalance the gender achievement gap at school, then biases in grades could be a mean to achieve a higher gender equality at school. These results also contribute to the ongoing debate on the use of grades as evaluation tools. The earlier teachers give grades to students, the higher the potential for discrimination. In several education systems, pupils do not receive any grades before they turn 11 or more (Sweden for instance). Again, the desirability of an early grading scheme depends on the goal assigned to grades. Finally, these results raise the question

of teachers' training on the impact of their stereotypes – conscious or unconscious - on their grades. French teachers have currently neither training nor information provided on the risks they face of judging their students through the lens of stereotypes. Making them aware of these risks might be a simple solution to significantly reduce biases in grades.

References

[1] Avvisati, Francesco, Marc Gurgand, Nina Guyon, and Eric Maurin. « Getting Parents Involved : a Field Experiment in Deprived Schools ». Review of Economic Studies 81, no 1(2014): 57-83.

[2] Bar, Talia, and Asaf Zussman. « Partisan grading ». American Economic Journal: Applied Economics 4, no 1 (2012): 30-48.

[3] Bénabou, Roland, and Jean Tirole. « Self-Confidence and Personal Motivation ». The Quarterly Journal of Economics 117, no 3 (8 janvier 2002): 871:915.

[4] Bertrand, Marianne, Esther Duflo, and Sendhil Mullainathan. « How Much Should We Trust Differences-in-Differences Estimates? ». National Bureau of Economic Research Working Paper Series No. 8841 (2002).

[5] Blank, Rebecca M. « The Effects of Double-Blind versus Single-Blind Reviewing: Experimental Evidence from The American Economic Review ». The American Economic Review 81, no 5 (1 décembre 1991): 1041-67.

[6] Bonesrønning, Hans. « The effect of grading practices on gender differences in academic performance ». Bulletin of Economic Research 60, no 3 (2008): 245-64.

[7] Bouguen, Adrien. «Adjusting content to every students' needs : further evidence from a teacher training program». Ongoing research.

[8] Breda, Thomas, and Son Thierry Ly. « Do professors really perpetuate the gender gap in science? Evidence from a natural experiment in a French higher education institution ». CEP WP, juin 2012.

[9] Burgess, Simon, and Ellen Greaves. Test Scores, Subjective Assessment and Stereotyping of Ethnic Minorities. The Centre for Market and Public Organisation. Department of Economics, University of Bristol, UK, septembre 2009.

[10] Crawford, Claire, Lorraine Dearden, and Costas Meghir. « When you are born matters: the impact of date of birth on child cognitive outcomes in England », octobre 2007.

[11] Dee, Thomas S. « A Teacher Like Me: Does Race, Ethnicity, or Gender Matter? ». American Economic Review 95, no 2 (2005): 158-65.

[12] ———. « Teachers and the Gender Gaps in Student Achievement ». The Journal of Human Resources 42, no 3 (1 juillet 2007): 528-54.

[13] Else-Quest, Nicole M, Janet Shibley Hyde, and Marcia C Linn. « Cross-national patterns of gender differences in mathematics: a meta-analysis ». Psychological Bulletin 136, no 1 (janvier 2010): 103-27.

[14] Falch, Torberg, and Linn Renée Naper. « Educational evaluation schemes and gender gaps in student achievement ». Economics of Education Review 36 (octobre 2013): 12-25.

[15] Fennema, Elizabeth, Penelope L. Peterson, Thomas P. Carpenter, and Cheryl A. Lubinski. « Teachers' Attributions and Beliefs about Girls, Boys, and Mathematics ». Educational Studies in Mathematics 21, no 1 (1 février 1990): 55-69.

[16] Goldin, Claudia. « Notes on Women and the Undergraduate Economics Major. » CSWEP Newsletter, Summer 2013, 15 édition.

[17] Goldin, Claudia, and Cecilia Rouse. « Orchestrating Impartiality: The Impact of "Blind" Auditions on Female Musicians ». American Economic Review 90, no 4 (septembre 2000): 715-41.

[18] Hinnerich, Björn Tyrefors, Erik Höglin, and Magnus Johannesson. « Are boys discriminated in Swedish high schools? ». Economics of Education Review 30, no 4 (août 2011): 682-90.

[19] Hoff, Karla, and Priyanka Pandey. « Discrimination, Social Identity, and Durable Inequalities ». The American Economic Review 96, no 2 (1 mai 2006): 206-11.

[20] Lavy, Victor. « Do gender stereotypes reduce girls' or boys' human capital outcomes? Evidence from a natural experiment ». Journal of Public Economics 92, no 10-11 (octobre 2008): 2083-2105.

[21] Lindahl, Erica. « Does gender and ethnic background matter when teachers set school grades? Evidence from Sweden ». Uppsala University Working paper (2007).

[22] Marsh, Herbert. W., and Rhonda G. Craven. « Academic self-concept: Beyond the dustbowl. » In G. D. Phye (Ed.), Handbook of classroom assessment. San Diego, CA: Academic Press., 1997, 131:198.

[23] Marsh, Herbert W., and Rhonda G. Craven. « The Pivotal Role of Frames of Reference in Academic Self-Concept Formation: The Big Fish-Little Pond Effect », 2001.

[24] Mechtenberg, Lydia. « Cheap Talk in the Classroom: How Biased Grading at School Explains Gender Differences in Achievements, Career Choices and Wages ». Review of Economic Studies 76, no 4 (2009): 1431-59.

[25] Neblett, Enrique W, Cheri L Philip, Courtney D Cogburn, and Robert M Sellers. « African American Adolescents' Discrimination Experiences and Academic Achievement: Racial Socialization as a Cultural Compensatory and Protective Factor ». Journal of Black Psychology 32, no 2 (5 janvier 2006): 199-218.

[26] OECD. PISA 2009 Results: What Students Know and Can Do – Student Performance in Reading, Mathematics and Science, 2010.

[27] Ouazad, Amine, and Lionel Page. « Students' Perceptions of Teacher Biases: Experimental Economics in Schools ». SSRN eLibrary, 1 juillet 2011.

[28] Ready, Douglas D., and David L. Wright. « Accuracy and Inaccuracy in Teachers' Perceptions of Young Children's Cognitive Abilities ». American Educational Research Journal 48, no 2 (1 avril 2011): 335-60.

[29] Robinson, Joseph Paul, and Sarah Theule Lubienski. « The Development of Gender Achievement Gaps in Mathematics and Reading During Elementary and Middle School Examining Direct Cognitive Assessments and Teacher Ratings ». American Educational Research Journal 48, no 2 (1 avril 2011): 268-302.

[30] Spencer, Steven J., Claude M. Steele, and Diane M. Quinn. « Stereotype Threat and Women's Math Performance ». Journal of Experimental Social Psychology 35, no 1 (janvier 1999): 4-28.

[31] Steele, Claude M., and Joshua Aronson. « Stereotype threat and the intellectual test performance of African Americans ». Journal of Personality and Social Psychology 69, no 5 (1995): 797-811.

[32] Tiedemann, Joachim. « Gender related beliefs of teachers in elementary school mathematics ». Educational Studies in Mathematics 41 (2000): 191-207.

[33] ———. « Teachers' Gender Stereotypes as Determinants of Teacher Perceptions in Elementary School Mathematics ». Educational Studies in Mathematics 50, no 1 (2002): 49-62.

[34] Trautwein, Ulrich, Oliver Ludtke, Herbert W. Marsh, Olaf Koller, et Jurgen Baumert. « Tracking, Grading, and Student Motivation: Using Group Composition and Status to Predict Self-Concept and Interest in Ninth-Grade Mathematics ». Journal of Educational Psychology 98, no 4 (novembre 2006): 788-806.

[35] Van Ewijk, Reyn. « Same Work, Lower Grade? Student Ethnicity and Teachers' Subjective Assessments ». Economics of Education Review 30, no 5 (octobre 2011): 1045-58.

[36] Wei, Thomas E. « Stereotype threat, gender, and math performance: evidence from the national assesment of educational progress ». Working Paper, Harvard University, 2009.

Table 1: **Balance check of the attrition**

	Sample with no missing	**Sample with Missing**	
	Mean	**Mean**	**Difference**
	(1)	**(2)**	**(3)=(2)-(1)**
Blind test scores			
FRA t1	.018	-.264	-0.283*
MAT t1	.017	-.234	-0.252*
Non-Blind test Scores			
FRA t1	.020	-.426	-0.447***
MAT t1	.030	-.321	-0.352**
Pupils' gender			
% Girls	.489	.389	-0.101***
Pupils' behavior			
% Disciplinary warning	.061	.076	0.016
% Excluded from class	.051	.079	0.028
% Temporary exclusion from school	.036	.018	-0.018
Parents' profession			
% Upper class	.182	.135	-0.047*
% Lower class	.698	.558	-0.140***
% Unemployed	.108	.171	0.062**
Number of observations	3964	555	

† Notes: Stars correspond to the following p-values: * p<.05; ** p<.01; *** p<.001. The full sample contains 4519 pupils. The reduced sample contains 3964 pupils for which all test scores are available during first term. 555 observations are considered as missing since one test score at least is missing during first term.

This table presents the differences between the reduced sample (with no test score missing) and the missing sample. The column "Difference" is the result of the regression of various dependant variables on a dummy indicating that the pupil has a score missing. All scores are standardized. Standard errors are robust and have been corrected for school-level clustering.

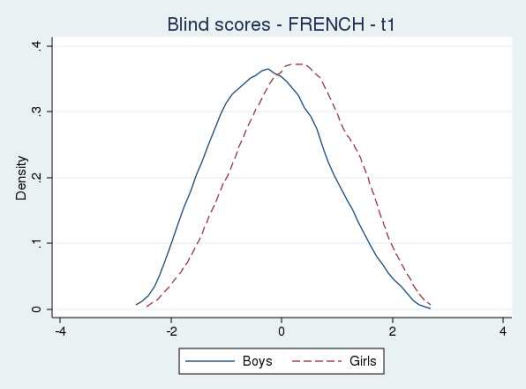
Parent's profession: Parents are considered to be upper class if they belong to the French administrative category "corporate manager" or "executive". Parents are classified as lower class if they belong to the categories "worker" or "white-collar worker". For both variables, the dummy takes the value "1" if at least one of the parents belongs to the category.

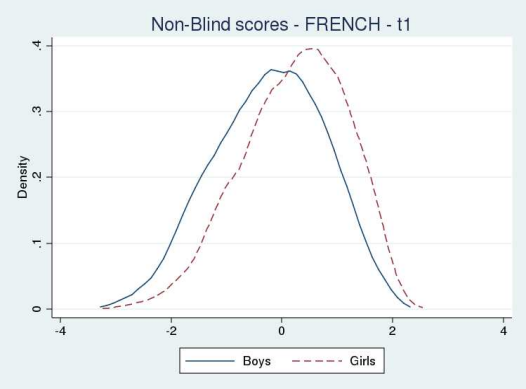
Table 2: **Comparison between boys' and girls' test scores**

	Term	**Subject**	**Girls** # obs	**Girls** Mean (1)	# obs	Mean (2)	**Difference between girls and boys mean scores** (3)=(1)-(2)	t-stat of the difference
Blind scores	T1	Mathematics	2020	-0.069	2127	0.078	**-0.147**	-4.74
		French	2022	0.228	2135	-0.205	**0.434**	14.32
		Both subjects	2041	0.077	2160	-0.070	**0.148**	5.24
	T3	Mathematics	1754	0.012	1804	0.053	**-0.040**	-1.22
		French	1761	0.227	1814	-0.167	**0.395**	12.16
		Both subjects	1863	0.107	1941	-0.083	**0.190**	6.67
Non-Blind Scores	T1	Mathematics	2042	0.098	2140	-0.070	**0.169**	5.53
		French	2024	0.236	2134	-0.224	**0.460**	15.24
		Both subjects	2060	0.168	2167	-0.153	**0.321**	11.63
	T3	Mathematics	2029	0.113	2127	-0.104	**0.217**	7.08
		French	2008	0.241	2104	-0.213	**0.454**	15.09
		Both subjects	2037	0.177	2134	-0.160	**0.338**	12.16

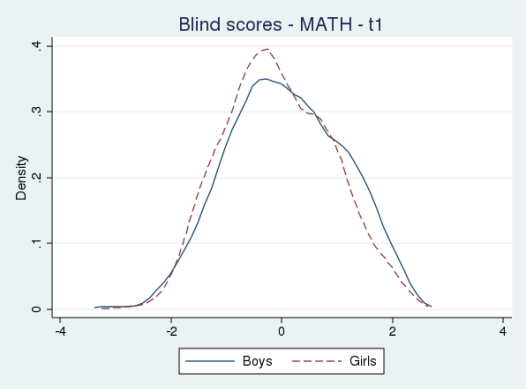
† Notes: All tests scores are standardized. Column (1) displays mean scores of girls in Mathematics and French, by nature of grading (blind scores at the top and non- blind scores at the bottom), and by term. Column (2) presents the same results for boys. Column (3) corresponds to the differences between girls' and boys' scores. All differences are highly statistically significant.

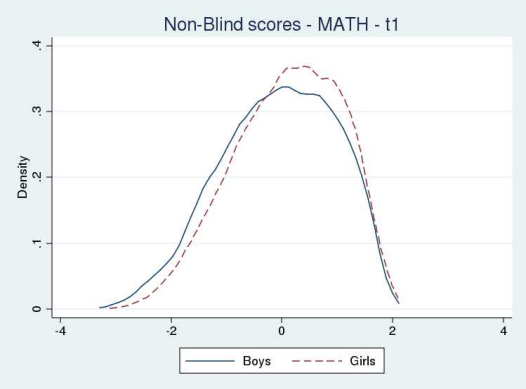
Graphics 1 and 2: Distribution of Blind and Non-Blind scores (first term) –**French**



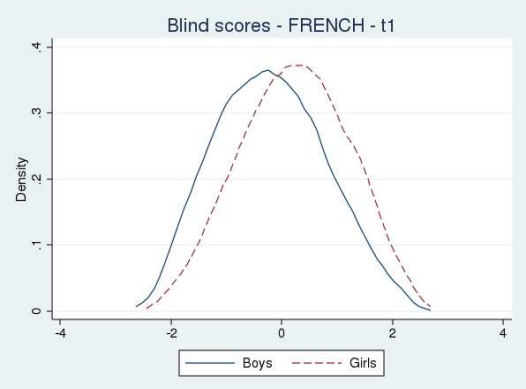


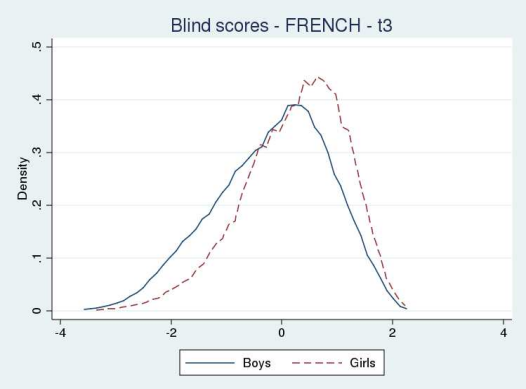
Graphics 3 and 4: Distribution of Blind and Non-Blind scores (first term) –**Math**



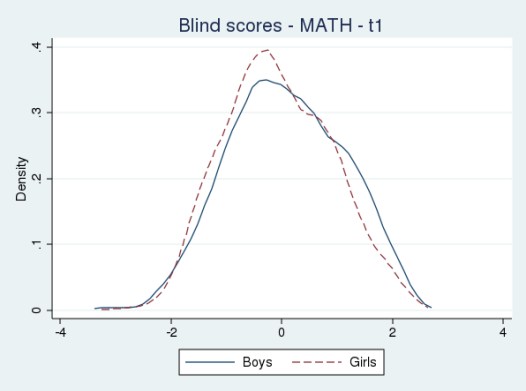


Graphics 5 and 6: Evolution of the distribution of Blind scores between T1 and T3 – **French**





Graphics 7 and 8: Evolution of the distribution of Blind scores between T1 and T3 – **Maths**



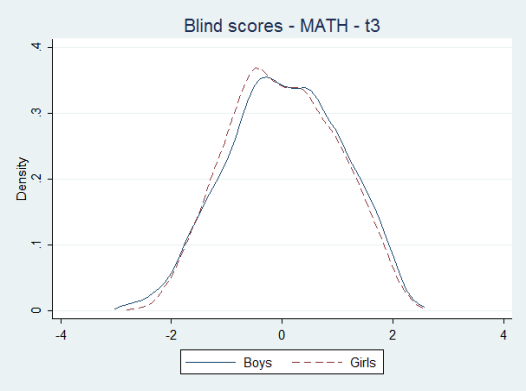


Table 3: **Estimation of the gender bias using Double-Differences**

	Reduced sample		**Full sample**	
	Math	**French**	**Math**	**French**
Dependent variable: Scores	(1)	(2)	(3)	(4)
Girl	-0.165***	0.412***	-0.153***	0.426***
	(0.028)	(0.019)	(0.028)	(0.018)
Non-Blind score	-0.148**	-0.024	-0.151**	-0.017
	(0.053)	(0.045)	(0.052)	(0.045)
Girl x Non-Blind	**0.323*** **	**0.043**	**0.318*** **	**0.027**
	(0.026)	**(0.031)**	**(0.027)**	**(0.032)**
Constant	0.097***	-0.191***	0.082**	-0.203***
	(0.025)	(0.025)	(0.025)	(0.024)
Class FE	yes	yes	yes	yes
Number of observations	8,136	8,116	8,329	8,315
R2	0.116	0.159	0.118	0.158

Notes: The dependent variable is the score (both blind and non-blind) obtained by a pupil in French or maths during the first term. Standard-errors are in parentheses and have been estimated with school level clusters. Stars correspond to the following p-values: * p<.05; ** p<.01; *** p<.001. All tests scores are standardized.
Each pupil has two observations: one for the blind score and one for the non-blind. The reduced sample contains 4068 pupils in math and 4058 in French for which both the blind and non-blind scores are non-missing. The full sample contains 4519 pupils. Some of them do not have two observations if the blind or non-blind score is missing.

Table 4: **Estimation of gender bias in math using Double-Differences with different control variables**

	Controls for punishments		**Controls for initial achievement**		
				1st decile	**10th decile**
Dependent variable: Scores	(1)	(2)	(3)	(4)	(5)
Girl	-0.148**	-0.212***	-0.153***	-0.134***	-0.098***
	(0.040)	(0.039)	(0.028)	(0.027)	(0.027)
Non-Blind score	-0.162*	-0.141	-0.151**	-0.186**	-0.105
	(0.065)	(0.067)	(0.052)	(0.054)	(0.055)
Girl x Non-Blind	**0.325*** **	**0.316*** **	**0.318*** **	**0.294*** **	**0.308*** **
	(0.030)	**(0.029)**	**(0.027)**	**(0.027)**	**(0.029)**
Punishment		-0.584***			
		(0.078)			
Punishment x Non-Blind		-0.141			
		(0.071)			
Punishment x Non-Blind x Girl		-0.293			
		(0.162)			
Decile				-1.631***	1.671***
				(0.039)	(0.028)
Decile x Non-Blind				0.261***	-0.386***
				(0.059)	(0.036)
Decile x Non-Blind x Girl				0.244***	-0.023
				(0.067)	(0.049)
Constant	0.070*	0.148***	0.082**	0.246***	-0.105***
	(0.031)	(0.033)	(0.025)	(0.026)	(0.027)
Class FE	yes	yes	yes	yes	yes
Number of observations	4,399	4,399	8,329	8,329	8,329
R2	0.103	0.135	0.118	0.313	0.304

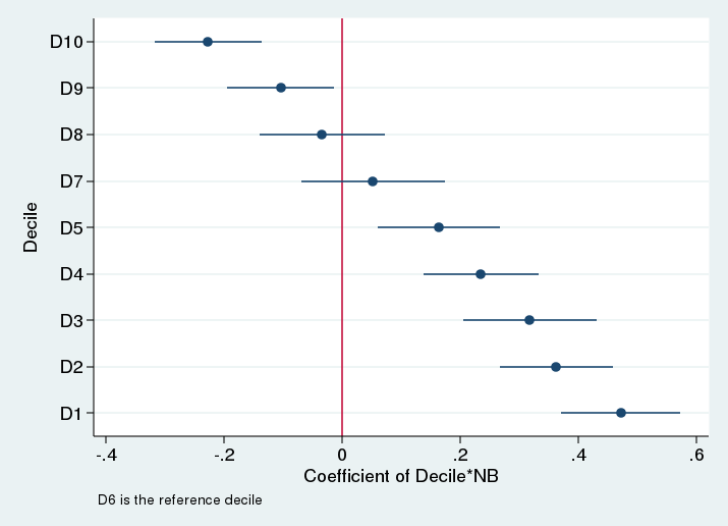
Notes: Standard-errors are in parentheses and have been estimated with school level clusters. Stars correspond to the following p-values: * p<.05; ** p<.01; *** p<.001. All tests scores are standardized. The dependent variable is the score (both blind and non-blind) obtained by a pupil in math during first term. The whole sample is used in columns 3 to 5. The sample used in columns 1 and 2 is the whole sample, to which pupils for which a punishment variable is missing have been removed.

Table 5: **Estimation of the gender bias with pupils' rank as dependant variable**

	Math	French
Dependent variable: Rank	(1)	(2)
Girl	1.193***	-2.806***
	(0.204)	(0.179)
Non-Blind score	1.289***	0.339**
	(0.101)	(0.110)
Girl x Non-Blind	-2.247***	-0.430*
	(0.177)	(0.175)
Constant	11.012***	12.973***
	(0.092)	(0.090)
Class FE	yes	yes
Number of observations	8,329	8,315
R2	0.048	0.091

Notes: The dependent variable is the score (both blind and non-blind) obtained by a pupil in math during first term. Standard-errors are in parentheses and have been estimated with school level clusters. Stars correspond to the following p-values: * p<.05; ** p<.01; *** p<.001. All tests scores are standardized.

Graphic 9: Decile distribution of the coefficient Decile*NB



Graphic 10: Correlation between month of birth and Blind score.

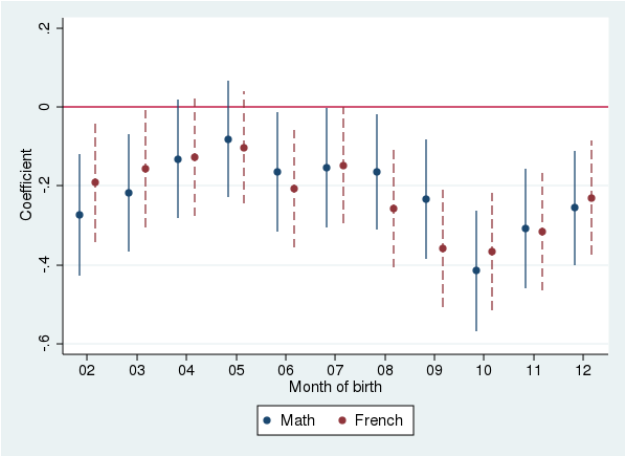


Table 6: **First stage - Correlation between Blind score and being born at the end of the year**

Dependant variable:	Math	French
Blind t1	(1)	(2)
Born End Year	-0.162***	-0.179***
	(0.044)	(0.041)
Girl	-0.176***	0.389***
	(0.042)	(0.041)
Punishment	-0.475***	-0.532***
	(0.073)	(0.068)
Grade retention	-0.336***	-0.210*
	(0.101)	(0.084)
Upper class	0.410***	0.411***
	(0.055)	(0.053)
Constant	0.124***	-0.157***
	(0.037)	(0.036)
Observations	2,168	2,120
R-squared	0.060	0.113
F stat	14.53	19.23

Notes: The dependent variable is the score (both blind and non-blind) obtained by a pupil in math during first term. Standard-errors are in parentheses and have been estimated with school level clusters. Stars correspond to the following p-values: * p<.05; ** p<.01; *** p<.001. All tests scores are standardized.

Table 7: **OLS and IV estimates of the reduced form**

	OLS		**IV**	
Dependant variable :	**Math**	**French**	**Math**	**French**
Non-Blind scores	(1)	(2)	(3)	(4)
Girl	0.262***	0.173***	**0.338*** **	**0.064**
	(0.028)	(0.043)	(0.032)	(0.059)
Blind score	0.753***	0.683***	1.082***	1.015***
	(0.019)	(0.030)	(0.096)	(0.108)
Constant	-0.093***	0.046	-0.036	0.262***
	(0.017)	(0.024)	(0.042)	(0.034)
Controls for sanctions	yes	yes	yes	yes
Class FE	yes	yes	yes	yes
Number of observations	2,168	2,120	2,168	2,120
R2	0.686	0.609	0.591	0.528

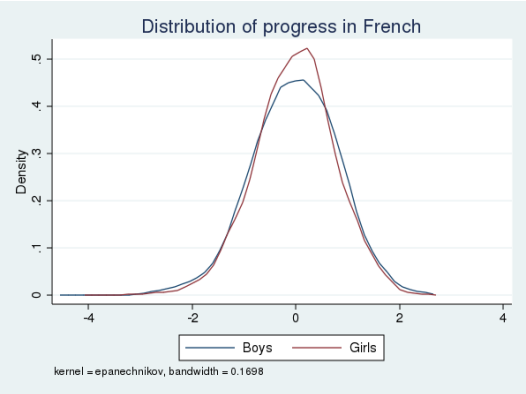
Notes: Standard-errors are in parentheses and have been estimated with school level clusters. Stars correspond to the following p-values : * p<.05; ** p<.01; *** p<.001. All tests scores are standardized. The unit of observation is a pupil. The sample contains 2168 pupils in math for which the blind score, non-blind score and punishment variable are non-missing, 2120 in French. The instrument is a dummy variable equal to one if a pupil is born between August and December in French, between September and December in Maths.

Table 8: **Comparison of DiD estimates of gender bias for first and last term**

	Math	French
	(1)	(2)
Coef Girl*Non-Blind	**0.318*** **	**0.027**
First term	**(0.027)**	**(0.032)**
Coef Girl*Non-Blind	0.251***	0.060
Last term	(0.034)	(0.042)

* p<.05; ** p<.01; *** p<.001. School level clustered s.e.
School level clustered s.e.
Full sample used.

Graphics 11 and 12: Comparison of boys and girls' progress - French and math



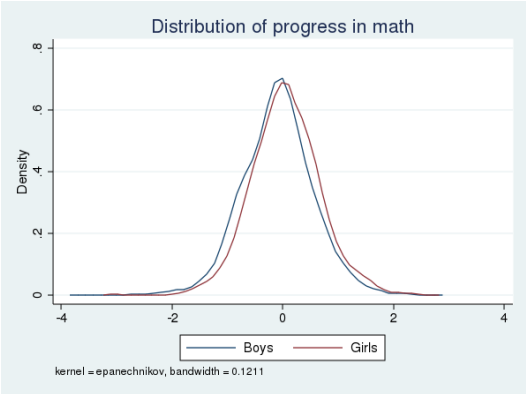
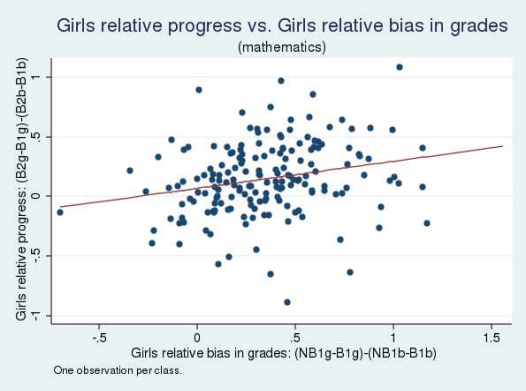


Table 9: **Comparison between boys' and girls' test scores**

		Term 1		**Term 3**		**Difference between t1 and t3**	t-stat of the
Subject	**Sex**	# obs	Mean	# obs	Mean	**mean scores**	difference
			(1)		(2)	(3)=(1)-(2)	
Mathematics	Boys	2127	0.078	1804	0.053	**0.025**	0.76
	Girls	2020	-0.069	1754	0.012	**-0,082**	-2.58
French	Boys	2135	-0.205	1814	-0.167	**-0.038**	-1.19
	Girls	2022	0.228	1761	0.227	**0.001**	0.01

† Note: All tests scores are standardized. Column (1) represents the mean blind score obtained by boys and girls during the first term. Column (2) presents mean blind scores during last term of the school year. Column (3) is the difference between the first term blind scores and the last term scores.

Graphics 13 and 14: Between-class comparison of girls' discrimination (relative to boys) and girls' progress (relative to boys).



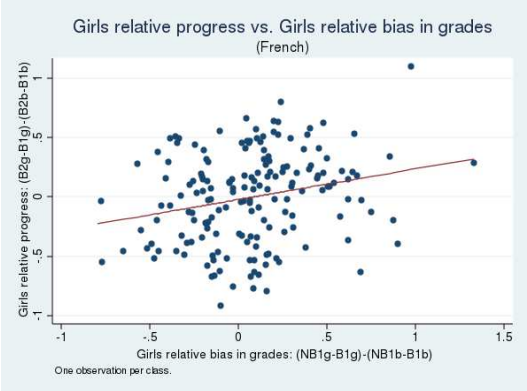


Table 10: **Effect of gender discrimination on girls' relative progress**

	Math	**French**
Dependent variable: $(B_{3G} - B_{3B})_c$	(1)	(2)
$[(NB_{1G} - B_{1G}) - (NB_{1B} - B_{1B})]_c$	0.287***	0.160
	(0.078)	(0.102)
$(B_{1G} - B_{1B})_c$	0.872***	0.874***
	(0.062)	(0.115)
Constant	-0.015	0.019
	(0.037)	(0.055)
Number of observations	175	171
R2	0.555	0.435

† Notes: The unit of observation is a class. The dependent variable is the gap between girls and boys third term blind score. Standard-errors are in parentheses and have been estimated with school level clusters. Stars correspond to the following p-values: * p<.05; ** p<.01; *** p<.001. All tests scores are standardized.

Appendix 1 : Omitted variable bias affecting ρ and α_2.

As stated in the section 3.4.1, the correlation between pupils' gender and the blind score implies that any bias on the coefficient ρ would also affect the estimate of the coefficient α_2. Using the formula of the omitted variable bias allows me to determine the direction of the bias that affects both ρ and α_2 (Bouguen, 2014).

The well-known formula of the omitted variable bias is :

$$E(b_1/X) = \beta_1 + (X_1'X_1)^{-1}X_1'X_2\beta_2 \tag{19}$$

where X_1 is a vector of the observed variables, X_2 is a vector of the unobserved variables, β_1 is the vector of the estimated coefficients of the observed variables, and β_2 is the vector of the coefficients of the unobserved variables.

In my setting, the observed variables are the blind score B_i and a pupil's gender G_i, and the unobserved variable is the error term affecting the blind score ϵ_{Bi} :

$$X_1 = \begin{pmatrix} B_1 & G_1 \\ B_2 & G_2 \\ \vdots & \vdots \\ B_n & G_n \end{pmatrix}, \beta_1 = \begin{pmatrix} \rho \\ \alpha_2 \end{pmatrix}, X_2 = \begin{pmatrix} \epsilon_{B,1} \\ \epsilon_{B,2} \\ \vdots \\ \epsilon_{B,n} \end{pmatrix}, \beta_2 = -\rho$$

Hence :

$$(X_1'X_1) = \begin{pmatrix} \sum_{i=1}^N B_i^2 & \sum_{i=1}^N B_iG_i \\ \sum_{i=1}^N B_iG_i & \sum_{i=1}^N G_i^2 \end{pmatrix}$$

$$(X_1'X_1)^{-1} = \frac{1}{\sum_{i=1}^N B_i^2 \sum_{i=1}^N G_i^2 - (\sum_{i=1}^N B_iG_i)^2} \begin{pmatrix} \sum_{i=1}^N G_i^2 & -\sum_{i=1}^N B_iG_i \\ -\sum_{i=1}^N B_iG_i & \sum_{i=1}^N B_i^2 \end{pmatrix}$$

$$(X_1'X_2) = \begin{pmatrix} \sum_{i=1}^N B_i\epsilon_{B_i} \\ \sum_{i=1}^N G_i\epsilon_{B_i} \end{pmatrix}$$

$$(X_1'X_1)^{-1}(X_1'X_2) = \frac{1}{\sum B_i^2 \sum G_i^2 - (\sum B_iG_i)^2} \begin{pmatrix} \sum G_i^2 \sum B_i\epsilon_{B_i} - \sum B_iG_i \sum G_i\epsilon_{B_i} \\ \sum B_i^2 \sum G_i\epsilon_{B_i} - \sum B_iG_i \sum B_i\epsilon_{B_i} \end{pmatrix}$$

To simplify notations,$\sum = \sum_{i=1}^N$

$$(X_1'X_1)^{-1}(X_1'X_2)\beta_2 = \frac{1}{\sum B_i^2 \sum G_i^2 - (\sum B_iG_i)^2} \begin{pmatrix} \rho\sum G_i^2 \sum B_i\epsilon_{B_i} - \rho\sum B_iG_i \sum G_i\epsilon_{B_i} \\ \rho\sum B_i^2 \sum G_i\epsilon_{B_i} - \rho\sum B_iG_i \sum B_i\epsilon_{B_i} \end{pmatrix}$$

The first row gives the bias which affects the estimates of the coefficient of B_i. The second row corresponds to the bias on the coefficient α_2 of the variable G_i :

$$\hat{\alpha_2} = \alpha_2 - \rho\frac{\sum B_iG_i \sum B_i\epsilon_{B_i}}{\sum B_i^2 \sum G_i^2 - (\sum B_iG_i)^2 - \sum B_i^2 \sum G_i\epsilon_{B_i}} \tag{20}$$

Dividing both the numerator and denominator by n, gives :

$$\hat{\alpha_2} = \alpha_2 - \rho \frac{V(B_i)Cov(G_i, \epsilon_{B_i}) - Cov(B_i, G_i)Cov(B_i, \epsilon_{B_i})}{V(B_i)[V(G_i) - \bar{G_i}] - Cov(B_i, G_i)^2} \tag{21}$$

Dividing both the numerator and denominator by $V(B_i)V(G_i)\sigma(\epsilon_{B_i})$ gives:

$$\hat{\alpha_2} = \alpha_2 - \rho \frac{r_{(G_i,\epsilon_{B_i})} - r_{(G_i,B_i)}r_{(B_i,\epsilon_{B_i})}}{1 + \frac{\bar{G_i}}{V(B_i)}r^2_{(B_i,G_i)}} \frac{\sigma(\epsilon_{B_i})}{\sigma(\epsilon_{G_i})} \tag{22}$$

where $\sigma(\epsilon_{B_i})$ is the standard deviation of ϵ_{B_i}, $\sigma(\epsilon_{G_i})$ is the standard deviation of ϵ_{G_i}, $r_{(G_i,\epsilon_{B_i})}$ is the correlation coefficient between G_i and ϵ_{B_i}...

Being a girl is assumed to be orthogonal to the shock affecting the blind score so that $r_{(G_i,\epsilon_{B_i})} = 0$. In a standard measurement-error model $r_{(B_i,\epsilon_{B_i})} = V(\epsilon_{B_i})$ so that we obtain:

$$\hat{\alpha_2} = \alpha_2 + \rho \frac{r_{(G_i,\epsilon_{B_i})}V(\epsilon_{B_i})}{1 + \frac{\bar{G_i}}{V(B_i)}r^2_{(B_i,G_i)}} \frac{\sigma(\epsilon_{B_i})}{\sigma(\epsilon_{G_i})} \tag{23}$$

Based on this formula, the direction of the bias depends on the sign of each of its elements: ρ is the correlation coefficient between the blind score and the non-blind score. It is positive. By definition, standard deviation and variances are also positive, as is the average value of the dummy G_i. Hence, the direction of the bias is fully determined by the sign of $r_{(G_i,\epsilon_{B_i})}$ which is positive in French and negative in mathematics. In the former subject girls perform better than boys for the standardized evaluation, while the opposite is observed in mathematics where girls perfom lower than boys on average.